

CLAIMS

What is claimed is:

1. A method for retrieving information using a search engine comprising the steps of:

- 5
- (a) retrieving a document to be indexed;
 - (b) generating a document extract corresponding to the document;
 - (c) decomposing the document extract into a plurality of tokens; and
 - (d) storing the plurality of tokens in a search index, wherein the search engine
- accesses the search index to retrieve information in one or more document extracts satisfying
- 10 a search query.

2. The method of claim 1, wherein the generating step (b) further comprises the steps of:

- 15
- (b1) extracting a portion of the document that characterizes the document's subject content to form the document extract; and
 - (b2) recording positional information of the portion extracted within the document.

3. The method of claim 2, further comprising the step of:

- 20
- (e) storing the document extract in a storage device.

4. The method of claim 3, wherein the storing step (d) further comprises:

(d1) storing the recorded positional information with the plurality of tokens.

5. The method of claim 4, wherein the extracting step (b1) further comprises the step of:

5 (b1i) extracting from the document a collection of sentences that are characteristic of the document's subject content to form a document summary.

6. The method of claim 4, wherein the decomposing step (c) further comprises:

10 (c1) selecting from the document extract one of a whole sentence, a portion of a sentence, a word, and a feature.

7. The method of claim 6, wherein the selecting step (c1) further comprises:

15 (c1i) selecting based on frequency of occurrence, word-salient-measure, proximity to the beginning of a paragraph, proximity to the beginning of the document, and proximity to or position within a heading or a caption.

8. The method of claim 1, wherein the document is a web-page in the Internet.

20 9. A computer readable medium containing programming instructions for retrieving information using a search engine comprising the instructions for:

- (a) retrieving a document to be indexed;
- (b) generating a document extract corresponding to the document;
- (c) decomposing the document extract into a plurality of tokens; and
- (d) storing the plurality of tokens in a search index, wherein the search engine

5 accesses the search index to retrieve information in one or more document extracts satisfying a search query.

10. The computer readable medium of claim 9, wherein the generating instruction
(b) further comprises the instructions for:

(b1) extracting a portion of the document that characterizes the document's
subject content to form the document extract; and

(b2) recording positional information of the portion extracted within the
document.

15 11. The computer readable medium of claim 3, further comprising the instruction
for:

(e) storing the document extract in a storage device.

20 12. The computer readable medium of claim 11, wherein the storing instruction
(d) further comprises the instruction for:

(d1) storing the recorded positional information with the plurality of
tokens.

13. The computer readable medium of claim 12, wherein the extracting instruction (b1) further comprises the instruction for:

(b1i) extracting from the document a collection of sentences that are characteristic of the document's subject content to form a document summary.

14. The computer readable medium of claim 12, wherein the decomposing instruction (c) further comprises the instruction for:

(c1) selecting from the document extract one of a whole sentence, a portion of a sentence, a word, and a feature.

15. The computer readable medium of claim 14, wherein the selecting instruction (c1) further comprises the instruction for:

(c1i) selecting based on frequency of occurrence, word-salient-measure, proximity to the beginning of a paragraph, proximity the beginning of the document, and proximity to and position within a heading and a caption.

16. The computer readable medium of claim 9, wherein the document is a web-page in the Internet.

17. A system for retrieving information, wherein the system includes a search

engine comprising:

means for retrieving a document from a document repository;

an information extractor coupled to the means for retrieving, wherein the information extractor generates a document extract corresponding to the document;

5 a storage device coupled to the information extractor for storing the document extract;

a search engine indexer coupled to the storage device for decomposing the document extract into a plurality of tokens; and

10 a search index coupled to the search engine indexer for storing the plurality of tokens, wherein the search engine accesses the search index to retrieve information in one or more document extracts satisfying a search query.

15 18. The system of claim 17, wherein the information extractor extracts a portion of the document that characterizes the document's subject content to form the document extract, and records positional information of the portion extracted within the document.

19. The system of claim 18, wherein the search index stores the positional information associated with the plurality of tokens.

20 20. The system of claim 19, wherein a token of the plurality of tokens comprises one of a whole sentence, a portion of a sentence, a word, and a feature of the document.

21. The system of claim 20, wherein the search engine indexer selects the plurality of tokens based on frequency of occurrence, word-salient-measure, proximity to the beginning of a paragraph, proximity the beginning of the document, and proximity to and position within a heading and a caption.

5

22. The system of claim 17, wherein the document respository is the Internet and the document is a web-page.

23. The system of claim 22, wherein the means for retrieving the document is a web crawler.

10